Journal of Applied Informatics and Computing (JAIC)
Vol.4, No.1, Juli 2020, pp. 84~88
e-ISSN: 2548-6861

Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining

Derisma
Sistem Komputer, Fakultas Teknologi Informasi
Universitas Andalas, Limau Manis Pauh Padang
derisma@fti.unand.ac.id

Article Info

Article history:

Received 2020-06-30
Revised 2020-07-11
Accepted 2020-07-13

Keyword:

Heart Disease,
Data Mining,
Algoritma Naive Bayes,
Random Forest,
Neural Network.

ABSTRACT

Heart disease is a disease that contributes to a relatively high mortality rate. The rate of human death caused by disease in the heart is a widespread problem in the world. The main objective of this study is to predict people with heart disease using the publicly available dataset in the UCI Repository with the Heart Disease dataset. To obtain the best classification algorithm is by comparing three Algoritma Naive Bayes, Random Forest, Neural Network algorithms, which are frequently used to predict people with heart disease. Comparison results show that Naive Bayes ' algorithm is a precise and accurate algorithm used to predict people with heart disease with a percentage of 83 %.

I. PENDAHULUAN

Penyakit jantung terdiri dari serangkaian gangguan yang mempengaruhi jantung. Ini termasuk masalah pembuluh darah seperti masalah detak jantung yang tidak teratur, otot jantung yang lemah, cacat jantung bawaan, penyakit kardio vaskular dan penyakit arteri koroner. Tingkat kematian oleh penyakit pada jantung termasuk kelompok tinggi di dunia. Bidang ilmu kedokteran sangat bergantung pada sarana otomatis berbasis komputer untuk diagnosis yang tepat dan akurat serta tepat waktu. Hal ini mengakibatkan pemeliharaan sejumlah besar data terkait pasien setiap hari. Data yang disimpan dapat digunakan sebagai sumber untuk memprediksi kemungkinan penyakit di masa depan yang membuat teknik penambangan data memainkan peran sentral untuk ekstraksi pengetahuan dan prediksi. Prediksi penyakit jantung adalah salah satu area yang tumbuh untuk prediksi tersebut. [1] [2]

Kekurangan Dokter, ahli dan mengabaikan gejala pasien menyebabkan tantangan besar yang dapat menyebabkan kematian, cacat bagi pasien. Oleh karena itu diperlukan sistem pakar yang berfungsi sebagai alat analisis untuk menemukan informasi dan pola tersembunyi dalam data medis penyakit dengar. Penambangan data adalah prosedur kognitif untuk menemukan pola pendekatan tersembunyi dari kumpulan data besar. Data besar yang tersedia dapat digunakan untuk mengekstrak informasi yang berguna dan menghubungkan semua atribut untuk membuat keputusan. Berbagai teknik dicantumkan dan diuji di sini untuk memahami tingkat akurasi masing-masing. Dalam studi sebelumnya, para peneliti menyatakan upaya mereka untuk menemukan model prediksi terbaik. Makalah ini mengusulkan sistem prediksi penyakit jantung. [3] [4]

Data Mining adalah teknik yang dilakukan pada basis data besar untuk mengekstraksi pola tersembunyi dengan menggunakan strategi kombinasional dari analisis statistik, pembelajaran mesin, dan teknologi basis data. Data mining medis adalah bidang penelitian yang sangat penting karena pentingnya dalam pengembangan berbagai aplikasi dalam domain perawatan kesehatan yang berkembang [5]. Data mining dikarakterisasi sebagai pencarian informasi yang berguna melalui kumpulan data yang sangat besar. Beberapa teknik kunci dan paling umum untuk penambangan data adalah aturan asosiasi, klasifikasi, pengelompokan, prediksi, dan model sekuensial. Untuk berbagai aplikasi, teknik penambangan data digunakan. Penambangan data memainkan peran penting dalam deteksi penyakit di industri perawatan kesehatan. Pasien harus diperlukan untuk mendeteksi sejumlah tes untuk penyakit ini. Namun, jumlah tes harus dikurangi dengan menggunakan teknik penambangan data [6].

Penelitian mengenai algoritma data mining telah banyak dilakukan dan dipublikasikan. Kaurdkk [7] melakukan perbandingan metode pengelompokan yang berbeda, algoritma Pohon Keputusan, Bayesian Network, dan Neural Network adalah tiga metode klasifikasi yang banyak digunakan berdasarkan perbandingan metode klasifikasi yang berbeda. Selain itu, model Decision Tree yang paling sering digunakan adalah CART dan C4.5, dan untuk mengevaluasi dan membandingkan model, akurasi prediksi banyak digunakan. Makalah ini merekomendasikan penggunaan dataset besar untuk menjamin kinerja model prediksi.

Lutimath dkk [3] melakukan Teknik Penambangan Data yang berbeda dapat digunakan untuk menganalisis masalah terkait jantung. Ada berbagai jenis Teknik Penambangan Data seperti Decision Tree, Naïve Bayesian, Support Vector Machine (SVM), pengklasifikasi K-NN, Pendekatan Hibrid, Jaringan JST Artificial Neural. Algoritma klasifikasi seperti Naïve Bayes (NB), Decision Tree (DT), dan Jaringan Syaraf Tiruan (JST) telah banyak digunakan untuk memprediksi penyakit jantung, di mana berbagai nilai akurasi diperoleh.

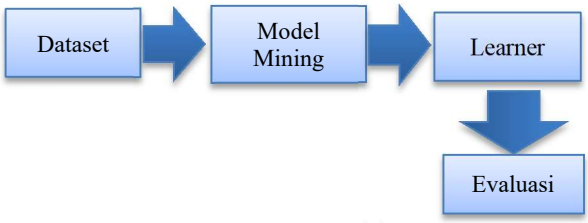
Penelitian tentang penerapan data mining untuk prediksi penyakit jantung juga telah ada dilakukan di Indonesia, Riani dkk [8] melakukan perhitungan menggunakan metode Data Mining dengan Algoritma Naive Bayes. Hasil dari penelitian ini mendapatkan akurasi sebesar 86% untuk 303 dataset yang yang diuji. Penelitian dengan memakai dataset penderita penyakit jantung menggunakan mengkomparasi 5 model yaitu decision tree, k-nearest neighbour, Naïve bayes, random forest, dan decision stump. Hasil penelitian diperoleh nilai akurasi tertinggi sebesar 80.38% pada algoritma random forest. Penelitian telah dilakukan Abdul Rohman [9] menggunakan algoritma neural network, neigboard knearest dan data pasien menggunakan C4.5 untuk akurasi prediksi penyakit penyakit jantung jantung dapat diamati bahwa metode terbaik adalah jaringan saraf untuk nilai akurarasi 86, 06 %. Wibisono dkk [10] membandingkan 4 algoritma Naïve Bayes, K-Nearest Neighbor, Decision Tree dan Random Forest, hasil klasifikasi dengan algoritma Random Forest memiliki rerata tingkat akurasi sebesar 85,668 % .

Selanjutnya penelitian ini membandingkan hasil algoritma terbaik dari penelitian sebelumnya yaitu Algoritma Naive Bayes, algoritma random forest, Neural Network, untuk mendapatkan algoritma paling baik dalam memperdiksi penyakit jantung. Tujuan utama dari penelitian ini adalah untuk mengidentifikasi algoritma klasifikasi paling baik dari penelitian terdahulu yang telah dilakukan oleh Riani dkk,. Abdul Rohman, dan Wibisono. Untuk itu dilakukan teknik data mining dengan membandingkan tiga algoritma klasifikasi data mining yaitu Algoritma Naive Bayes, Random Forest, Neural Network.

II. METODE

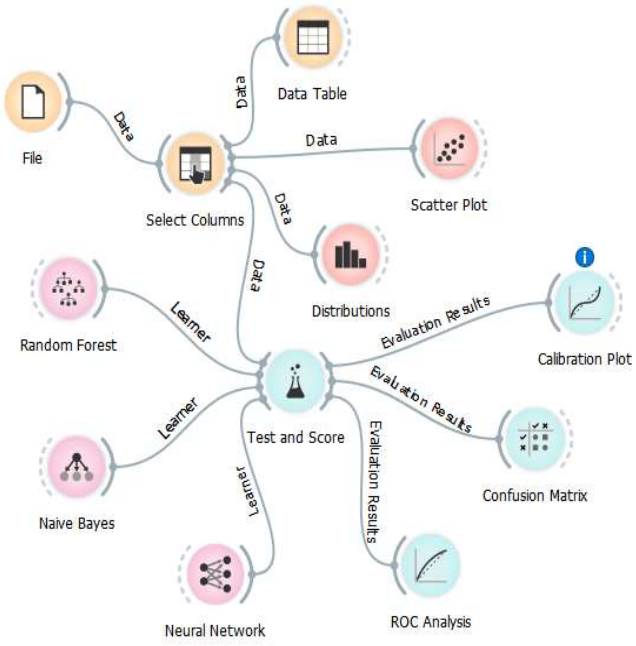
Tahapan penelitian ini dijelaskan sebagai berikut.

1) Pengumpulan dataset, yang mana dataset yang akan diuji dalam penelitian ini yaitu Heart Disease dataset dari UCI Repository.
2) Study literatur, study literatur ini merupakan pengumpulan artikel-artikel/ jurnal-jurnal penelitian yang terkait dengan penilitian yang akan diteliti dalam makalah ini, yakni terkait dengan prediksi penyakit jantung.
3) Pemilihan model penelitian. Dalam makalah ini akan menggukan model terbaik yang telah dilakukan oleh beberapa peneliti sebelumnya, Algoritma Naive Bayes, Random Forest, Neural Network
4) Learner atau pelatihan dari dataset
5) Evaluasi Prediksi dengan AUC, CA, F1, Precision, Recall, Confusion Matrix dan ROC analysis
6) Analisa dan kesimpulan, yakni menganalisi hasil pengujian yang dilakukan dan menyimpulkan analisis tersebut.



Gambar 1. Proses Data Mining

Adapun untuk proses data mining disini menggunakan tool data mining yaitu orange python dengan *workflow* sebagai berikut.



Gambar 2.Workflow Proses Prediksi Penyakit Jantung

Beberapa *performance metrics* yang umum dan juga sering digunakan, yaitu sebagai berikut.

1) Akurasi

Akurasi dapat diilustrasikan seberapa akurat model untuk mengklasifikasikan dengan benar. Dengan demikian, akurasi prediksi adalah perbandingan jumlah data positif benar dan negatif benar dengan data keseluruhan. Dengan kata lain, akurasi merupakan tingkat kedekatan nilai dari prediksi dengan nilai yang sebenarnya. Nilai dari akurasi dapat dilihat pada persamaan (1).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(1)}$$

2) Presisi

Presisi dapat digambarkan sebagai tingkat akurasi data yang diminta dengan hasil prediksi yang diberikan oleh model. Maka dapat diartikan bahwa presisi itu merupakan perbandingan dari prediksi benar positif dibagi dengan keseluruhan hasil yang diprediksi positif. Dengan kata lain dari semua kelas positif yang telah diprediksi dengan benar, berapa banyak data yang benar-benar positif. Nilai dari presisi dapat dilihat pada persamaan (2).

$$Presisi = \frac{TP}{TP+FP} \quad \text{.....(2)}$$

3) Recall

Recall atau sensitifitas dapat digambarkan sebagai keberhasilan model dalam mendapatkan sebuah informasi. Maka dapat dikatakan bahwa *recall* itu adalah rasio prediksi benar positif dibagi dengan keseluruhan data yang benar positif. Nilai dari *recall* dapat dilihat pada persamaan (3).

$$Presisi = \frac{TP}{TP+FN} \quad \text{.....(3)}$$

4) Specificity

Specificity dapat digambarkan dengan ketepatan memprediksi negatif dibagi dengan keseluruhan data negatif. Nilai dari *specificity* dapat dilihat pada persamaan (4)

$$Specificity = \frac{TN}{TN+F} \quad \text{.....(4)}$$

5) Skor F1

Skor F1 dapat dikatakan sebagai perbandingan rata-rata presisi dan *recall* yang dibobotkan. Skor F1 dapat dikatakan terbaik jika ada semacam keseimbangan antara presisi dan *recall* dalam sistem. Nilai dari skor F1 dapat dilihat pada persamaan (5).

$$Skor\ F1 = 2x\frac{Recall\ x\ Presisi}{Recall+Pres} \quad \text{.....(5)}$$

III. HASIL DAN PEMBAHASAN

Metode yang dipakai dalam penelitian ini adalah algoritma kalisifikasi dengan Algoritma Naive Bayes, Random Forest, dan Neural Network yang digunakan untuk mengembangkan sistem prediksi untuk menganalisis dan memprediksi kemungkinan penyakit jantung. Untuk mendapatkan kinerja algoritma klasifikasi ini menggunakan dataset tersedia umum di UCI Repository dengan dataset Penyakit Jantung. Dataset yang digunakan memiliki 14 atribut terdiri dari data nominal dan numerik. Kelas data sasaran adalah tidak adanya penyakit jantung dan adanya penyakit jantung. Dapat dilihat pada gambar di bawah.

Info

Heart Disease dataset
Data on the presence of heart disease in patients.

303 instance(s)
13 feature(s) (0.2% missing values)
Classification; categorical class with 2 values (no missing values)
0 meta attribute(s)

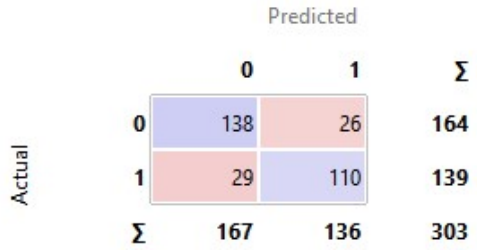
Columns (Double click to edit)

	Name	Type	Role
1	age	N numeric	feature
2	gender	C categorical	feature
3	chest pain	C categorical	feature
4	rest SBP	N numeric	feature
5	cholesterol	N numeric	feature
6	fasting blood sugar > 120	C categorical	feature
7	rest ECG	C categorical	feature
8	max HR	N numeric	feature
9	exerc ind ang	C categorical	feature
10	ST by exercise	N numeric	feature
11	slope peak exc ST	C categorical	feature
12	major vessels colored	N numeric	feature
13	thal	C categorical	feature
14	diameter narrowing	C categorical	target

Gambar 3. Heart Disease Dataset

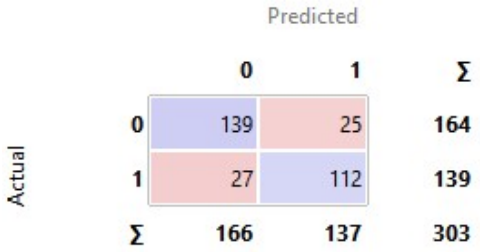
Sistem ini terdiri dari beberapa tahap yaitu preprocessing untuk menghilangkan data– data yang noisy ataupun redudansi, selanjutnya melakukan klasifikasi dengan Algoritma Naive Bayes, Random Forest, dan Neural Network. Setelah itu dilakukan validasi untuk melihat kinerja algoritma yaitu AUC, CA, F1, Precission, dan Recall dari masing–masing algoritma guna mendapatkan algoritma

dengan akurasi terbaik. Melakukan percobaan ini dengan menggunakan aplikasi Orange Python dengan sebuah daset yang memiliki 14 ribut yang terdiri dari data numerik dan nominal. Algoritma klasifikasi yang diuji yaitu Algoritma Naive Bayes, Random Forest, dan Neural Network. Nilai Confusion matrix menghitung akurasi pada konsep data mining atau sistem pendukung keputusan, berfungsi untuk melakukan analisis apakah classifier tersebut baik dalam mengenali kelas yang berbeda.Berikut merupakan confusion matrix dari Algoritma Naive Bayes, Random Forest, dan Neural Network.



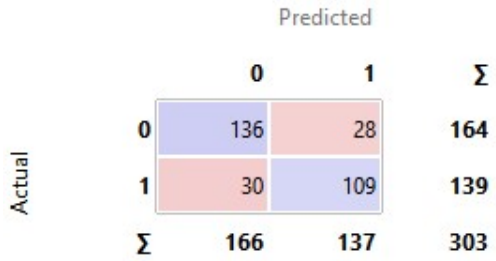
Gambar 4. Confusion Matrix Random Forest

Algoritma Random Forest pada Gambar 4 terdapat 138+110 prediksi yang benar dan 29+26 prediksi yang salah dari total 303 data testing. Sehingga, tingkat akurasi algoritma (precision) ini dapat dihitung 248/303*100% = 82 %.



Gambar 5. Confusion Matrix Naive Bayes

Algoritma Naïve Bayes pada Gambar 5 terdapat 139+112 prediksi yang benar dan 25+27 prediksi yang salah dari total 303 data testing. Sehingga, tingkat akurasi algoritma (precision) ini dapat dihitung 251/303*100% = 83 %.



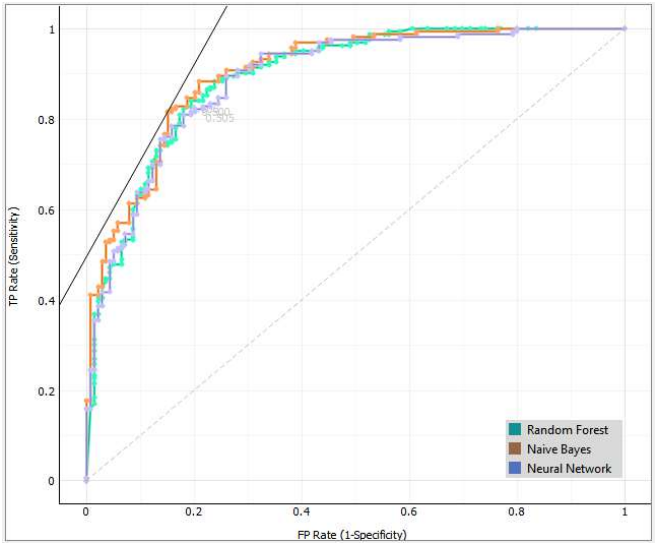
Gambar 6. Confusion Matrix Neural Network

Algoritma Neural Network pada Gambar 6 terdapat 136+109 prediksi yang benar dan 30+28 prediksi yang salah dari total 303 data testing. Sehingga, tingkat akurasi algoritma (precision) ini dapat dihitung 245/303*100% = 81 %.

TABEL 1
KINERJA ALGORITMA

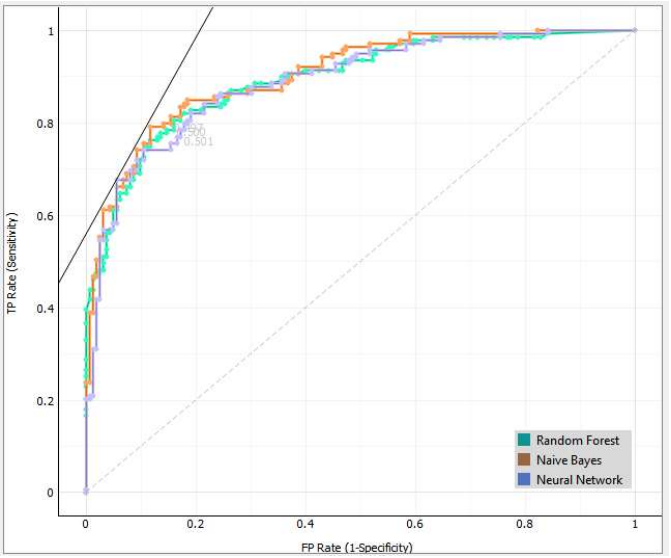
Model	AUC	CA	F1	Precision	Recall
Naive Bayes	0.90	0.83	0.83	0.83	0.83
Neural Network	0.89	0.81	0.81	0.81	0.81
Random Forest	0.89	0.82	0.82	0.82	0.82

Pada Tabel 1 dapat diketahui nilainya berdasarkan rumus dari Tabel 1. AUC, CA, F1, Precission, dan Recall yaitu metode yang dipakai pada saat pengujian data. CA dapat berfungsi untuk akurasi dari dataset yang dipilih. Precission adalah akurasi data yang memungkinkan dua kejadian yaitu 1 dan 0. Recall berfungsi untuk mengukur rasio. F1 yaitu perbandingan anatara recall dan presisi. AUC digunakan untuk mewakili probabilitas.



Gambar 7. ROC analysis Target Class 0

Pada gambar 7, ROC analysis masing-masing metode ditandai dengan warna grafik yang berbeda. Untuk Random Forest ditandai dengan warna hijau, Naïve bayes dengan warna orange, Neural Network ditandai dengan biru. Pada ketiga metode dalam satu grafik terlihat bahwa nilai sensitivity dan specificity tidak lebih dari 1 dan ketika nilai target class diubah, pada semua metode tidak nampak secara jelas perubahan pada sensitifity dan specificity.



Gambar 8. ROC analysis Target Class 1

Dari analysis kurva ROC Gambar 7 dan Gambar 8 dapat dilihat kinerja algoritma klasifikasi. Semakin dekat kurva mengikuti batas kiri dan kemudian batas atas ruang ROC, semakin akurat classifier tersebut. Dari ROC Analysis tersebut dapat diketahui bahwa model Naïve Bayes memiliki keakuratan dari classifier lebih baik dibandingkan dari model Neural Network dan Random Forest. Hal ini terlihat pada setiap class dapat dilihat bahwa kurva dari Naïve Bayes rata-rata mendekati sumbu axis atau sumbu Y yang menandai bahwa Naïve Bayes memiliki keakuratan classifier yang optimal.

IV. KESIMPULAN

Penelitian dengan menggunakan dataset penderita penyakit jantung dengan mengkomparasi 3 algoritma Algoritma Naive Bayes, Random Forest, Neural Network. Dengan menggunakan data mining pada aplikasi orange bersama dengan metode eksplorasi lainnya. Dataset terdiri 14 atribut terdiri dari data nominal dan numerik, target kelas data set ini adalah tidak adanya (0) dan adanya penyakit jantung (1). Hasil perbandingan menunjukkan bahwa dalam penggunaan algoritma klasifikasi data mining yang digunakan yaitu Algoritma Naive Bayes, Random Forest, Neural Network dapat kita lihat bahwa algoritma Naive Bayes adalah algoritma yang tepat dan akurat digunakan untuk dapat melakukan prediksi penderita penyakit jantung dengan persentase sebesar 83 %.

DAFTAR PUSTAKA

[1] R. Alizadehsani, M. Roshanzamir, M. Abdar, A. Beykikhoshk, A. Khosravi, M. Panahiazar, A. Koohestani, F. Khozeimeh, S. Nahavandi and N. Sarrafzadegan, "A database for using machine learning and data mining techniques for coronary artery disease diagnosis," *Scientific data,* vol. 6, no. 1, p. 227, 23 10 2019.

[2] S. Aydin, M. Ahanpanjeh and S. Mohabbatiyan, "Comparison and Evaluation Data Mining Techniques in the Diagnosis of Heart Disease," *International Journal on Computational Science & Applications,* vol. 6, no. 1, pp. 1-15, 29 2 2016.

[3] N. M. Lutimath, C. Chethan and B. S. Pol, "Prediction of heart disease using machine learning," *International Journal of Recent Technology and Engineering,* vol. 8, no. 2 Special Issue 10, pp. 474-477, 1 9 2019.

[4] L. Yahaya, N. David Oye and E. J. Garba, "A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques," *American Journal of Artificial Intelligence,* vol. 4, no. 1, pp. 20-29, 2020.

[5] R. Ghorbani and R. Ghousi, *Predictive data mining approaches in medical diagnosis: A review of some diseases prediction,* vol. 3, Growing Science, 2019, pp. 47-70.

[6] M. Tarawneh, O. E. Hct and U. Alfujairah, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques," 2019.

[7] A. Kaur, " Heart Disease Prediction Using Data Mining Techniques: A SURVEY," *International Journal of Advanced Research in Computer Science,* vol. 9, no. 2, pp. 569-572, 20 2 2018.

[8] A. Riani, Y. Susianto and N. Rahman, "Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Mengunakan Metode Naive Bayes," *Journal of Innovation Information Technology and Application (JINITA),* vol. 1, no. 01, pp. 25-34, 26 12 2019.

[9] A. Rohman and J. Banjarsari Barat No, "Komporasi Metode Klasifikasi Data Mining Untuk Prediksi Penyakit Jantung," 2016.

[10] A. B. Wibisono and A. Fahrurozi, " Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner," *Jurnal Ilmiah Teknologi dan Rekayasa,* vol. 24, no. 3, pp. 161-170, 2019.